J. Dairy Sci. 96:4666–4677
http://dx.doi.org/10.3168/jds.2012-6316


# Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle

**P. Ma,*†[1] R. F. Brøndum,* Q. Zhang,† M. S. Lund,* and G. Su*[1]**
*Centre for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK 8830 Tjele, Denmark
†Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture, Department of Animal Genetics and Breeding, College of Animal Science and Technology, China Agricultural University, China 100193

## ABSTRACT

This study investigated the imputation accuracy of different methods, considering both the minor allele frequency and relatedness between individuals in the reference and test data sets. Two data sets from the combined population of Swedish and Finnish Red Cattle were used to test the influence of these factors on the accuracy of imputation. Data set 1 consisted of 2,931 reference bulls and 971 test bulls, and was used for validation of imputation from 3,000 markers (3K) to 54,000 markers (54K). Data set 2 contained 341 bulls in the reference set and 117 in the test set, and was used for validation of imputation from 54K to high density [777,000 markers (777K)]. Both test sets were divided into 4 groups according to their relationship to the reference population. Five imputation methods (Beagle, IMPUTE2, findhap, AlphaImpute, and FImpute) were used in this study. Imputation accuracy was measured as the allele correct rate and correlation between imputed and true genotypes. Results demonstrated that the accuracy was lower when imputing from 3K to 54K than from 54K to 777K. Using various imputation methods, the allele correct rates varied from 93.5 to 97.1% when imputing from 3K to 54K, and from 97.1 to 99.3% when imputing from 54K to 777K; IMPUTE2 and Beagle resulted in higher accuracies and were more robust under various conditions than the other 3 methods when imputing from 3K to 54K. The accuracy of imputation using FImpute was similar to those results from Beagle and IMPUTE2 when imputing from 54K to high density, and higher than the remaining 2 methods. The results also showed that a closer relationship between test set and reference set led to a higher accuracy for all the methods. In addition, the correct rate was higher when the minor allele frequency was lower, whereas the correlation coefficient was lower when the minor allele frequency was lower. The results indicate that Beagle and IMPUTE2 provide the most robust and accurate imputation accuracies, but considering computing time and memory usage, FImpute is another alternative method.

**Key words:** imputation, relationship, minor allele frequency

## INTRODUCTION

Analyses based on genomic data such as genomic selection (**GS**) and genome-wide association study (**GWAS**) has been widely used in cattle breeding. Both GS and GWAS require a large number of individuals to be genotyped with a large number of markers spread along the genome such as SNP markers (Meuwissen et al., 2001; MacLeod et al., 2010). One of the factors affecting the accuracy of genomic prediction and GWAS is the density of SNP markers (Habier et al., 2009; Meuwissen, 2009). In principle, higher density should lead to better prediction and more accurate QTL mapping, because of stronger linkage disequilibrium (**LD**) between markers and causative mutations. However, higher density of markers also means higher cost of genotyping.

Currently, the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA; Matukumalli et al., 2009) has been widely used for genomic prediction in dairy cattle (Hayes et al., 2009; Su et al., 2010; VanRaden and Sullivan, 2010; Lund et al., 2011). However, some countries have genotyped several bulls with the 777,000-marker (**777K**; high-density, **HD**) chip with the intention of increasing the accuracy of genomic prediction, especially for genomic prediction across breeds. Higher density increases the persistence of the LD phase among populations, which is more beneficial for genomic prediction across populations than within population (Su et al., 2012). In addition, it has been proposed to genotype more individuals with a low-density chip (e.g., Bovine3K with 2,900 markers or BovineLD with 6,909 markers; Illumina Inc.) to increase the selection intensity at low cost (Boichard et al., 2012; Wiggans et al., 2012). When different SNP chips are used

Received October 28, 2012.
Accepted March 28, 2013.
[1]Corresponding authors: peipei.ma@agrsci.dk and guosheng.su@agrsci.dk

in genomic selection, imputation of missing genotypes is an important approach to make efficient use of all available marker data. Imputation is also necessary for GS or GWAS using joint reference data from exchange of genotypes between countries, where different chips are used for genotyping animals. Even for marker data from the same chip, imputation is useful for increasing the call rate of genotyped animals.

Various methods have been developed for imputation of missing genotypes. Some methods use pedigree information, whereas others do not. For example, AlphaImpute (Hickey et al., 2012b), FImpute (Sargolzaei et al., 2011), and findhap (VanRaden et al., 2011) use pedigree information, although pedigree information is not compulsory for FImpute. These methods were developed for animals and plants, as they can efficiently use complex pedigrees. Beagle (Browning and Browning, 2009) and IMPUTE2 (Howie et al., 2009), which were developed for human genetics, usually do not use pedigree information for imputation of marker data of livestock.

Imputation of missing marker genotypes is based on available marker data from a given population. The population structure and frequencies of marker genotypes in the given population have an influence on the imputation accuracy (Druet et al., 2010; Dassonneville et al., 2011; Hickey et al., 2012a). Because of differences in algorithms and different uses of information sources, the superiority of various imputation methods may differ in different imputation scenarios. Therefore, it is necessary to find the optimal imputation method and strategy to be used in the population of interest.

It has been reported that genetic variants with low frequency play a very important role in complex traits and may have larger effects than the common variants (Manolio et al., 2009). Therefore, it is necessary to investigate the efficiency of imputing markers with low minor allele frequency (**MAF**).

Although several studies have already been done on imputation of missing genotypes, most of these studies dealt with imputation methods and relationships between genotyped animals for imputation from a low-density panel to the 54,000-marker (**54K**) panel. It is necessary to compare imputation accuracy in relation to imputation methods, relationship between genotyped animals, marker densities, and marker MAF simultaneously in a given population. A simultaneous comparison is important for assessing the effect of each single factor and the combined effect of many factors on the accuracy of imputation.

The objectives of this study were 4-fold: (1) validating the accuracy of imputation from 3,000 markers (**3K**) to 54K and from 54K to HD using different methods in a combined population of Swedish and Finnish Red Cattle, (2) exploring the effect of the relationship between reference and test sets on imputation accuracy, (3) comparing the sensitivity of different imputation methods with the relatedness between reference and test population, and (4) investigating the efficiency of imputation for markers with low MAF.

## MATERIALS AND METHODS

### *Data*

Two data sets composed of bulls from Swedish and Finnish Red populations were used to validate imputation procedures in this study. These 2 populations have strong genetic links due to some bulls in common use (Brøndum et al., 2011). Data set 1 consisted of 3,902 bulls (born between 1960 and 2006) genotyped with the Illumina BovineSNP50 BeadChip (54K). There were 3,893 animals with both parents, 7 animals without dam, and 2 animals without any parent in the pedigree. Data set 2 contained a subset of data set 1, with 458 bulls (born between 1960 and 2005) that were genotyped with both the Illumina BovineHD BeadChip (777K; HD; Illumina Inc.) and the 54K chip. In this data set, 450 animals had both parents, 6 animals had no dam, and 2 animals had neither sire nor dam in the pedigree.

Two imputation scenarios with regard to marker density were investigated in this study. One was imputation from 3K to 54K, and the other was imputation from 54K to HD. In the validation of imputation from 3K to 54K, bulls in data set 1 were divided into a reference population and a test population by birth date so that reference bulls were born before October 1, 2001. For the test population, 3K marker data were derived from the 54K data by masking the markers that were not on the Illumina Bovine3K BeadChip. In the validation of imputation from 54K to HD, bulls in data set 2 were divided into a reference and a test population. The test population comprised 117 bulls born after April 1, 1999, and their 54K marker genotypes were used as test data. Furthermore, markers that were in the 54K chip but not in the HD map were excluded from the test data.

For all data sets, monomorphic markers were deleted. Minor allele frequencies and deviation from Hardy-Weinberg equilibrium were not used for editing marker data because markers with low MAF and deviation from Hardy-Weinberg equilibrium may be meaningful in genomic selection and GWAS, and one of our objectives was to compare imputation accuracy for the markers with different MAF. In addition, markers on the X chromosome were excluded, because no link between sire and son for markers on the X chromosome